Contents

SI-1. Library Curation

a. R-BIND

The RNA-targeted BIoactive ligaNd Database (R-BIND) is inclusive of organic small molecule probes that were reported in the literature through December 2016. The ligands were included if they satisfied the following criteria: i) highlighted by the author in the conclusion; ii) had activity in cell culture and/or animal models; iii) had evidence of binding to the target *in vitro*; and iv) had a molecular weight < 2,000 amu. Three ligands (R-BIND (SM) 0034, R-BIND (SM) 0037, and R-BIND (MV) 0034) were not reported to bind to the target *in vitro*; however, binding was reported for similar ligands in the same reference. The molecular weight cutoff was chosen because the majority of FDA-approved chemical entities (> 99.5%) are < 2,000 amu. The database (n = 104) was divided into two sub-libraries: monovalent small molecules (R-BIND (SM), n = 67) and multivalent ligands (R-BIND (MV), n = 37). A complete list of compounds can be found in the accompanying SI Excel file (RBIND.xls).

In general, the MV sub-library was segregated from the SM sub-library based on the presence of an alkyl, aryl, or peptide-like linker between multiple binding moieties and a molecular weight of > 500 amu. There were three exceptions to the molecular weight cutoff based on author descriptions: R-BIND (SM) 0025 (590 amu), R-BIND (MV) 0005 (496 amu), and R-BIND (MV) 0007 (496 amu). R-BIND 0025 was identified by Inforna to bind a 1 X 1 internal loop. R-BIND (MV) 0005 and 0007 were described as containing three binding moieties in the publication.

Small molecules targeting ribosomal RNA (rRNA) were excluded from R-BIND. rRNA is decidedly unique compared to other RNAs as it is both an active catalyst and highly abundant.[1] rRNA constitutes 80-85% of cellular RNA by mass, followed by tRNA (10-13%), mRNA (3-5%), and other non-coding RNAs (< 2%).[2] It has been proposed that these disparities could lead to distinct specificity requirements and thus distinct small molecule properties. Important differences between ribosomal and non-ribosomal RNA targeting small molecules have been reported.[3]

b. NALDB

RNA-binding molecules from the Nucleic Acid Ligand Database (NALDB)[4] were collected in January 2017 from the following website sections: Double-stranded RNA binding ligands, G-quadruplex RNA binding ligands, nucleic acid aptamer binding ligands, and nucleic acid special structure binding ligands. DNA ligands were removed from the latter two subsets. If the binding detail was ambiguous or not listed, the reported reference(s) were checked for evidence of RNA-binding and “no binding” entries were removed (32). Molecule SMILES were checked for accuracy and duplicate ligands were removed if present. Additionally, NALDB entries were excluded if they contained any of the following: < 3 carbon atoms (2), > 2000 amu (17), bound to DNA/RNA hybrid structures (6), contained a metal complex (1), or were already present in the R-BIND (16). The final NALDB library contained a total of 306 members.

To accurately compare the NALDB and R-BIND, the library was filtered to remove aminoglycosides (n = 71), which were identified by the presence of glycosidic linkages and/or the designations within the references from the NALDB website. Non-aminoglycoside molecules that were reported to bind to the ribosome (n = 19) were also removed.

The remaining 192 molecules were divided into the following categories, which were used for analysis: monovalent small molecules (NALDB (SM), n = 173), and multivalent ligands (NALDB (MV), n = 44). Multivalent (MV) ligands were differentiated from the remaining small molecules (SM) by the presence of multiple or repeating binding moieties and are generally characterized by a molecular weight of > 500 amu. There were 25 ligands classified as SM with a molecular weight > 500 amu. These small molecules are g-quadraplex binding ligands (n = 16), larger natural products (n = 1) or dyes (n = 3), or identified in the NALDB listed reference as single RNA module or monomer (n = 5). There were 3 ligands classified as MV with a molecular weight < 500 amu. The references listed in the NALDB for these ligands used the term two units or dimer. A complete list of compounds can be found in the accompanying SI Excel file (OtherLibraries.xls).

We also removed small molecules with reported biological activity from the NALDB libraries by comparing to R-BIND. It cannot be determined whether the remaining molecules were tested in cell culture and/or animal models and were unsuccessful or if the experiments were not conducted. Furthermore, some of the NALDB (SM) and (MV) libraries were tested for binding to aptamers or secondary structures, which often cannot be directly tested in cell culture or animal models. We emphasize that these libraries serve only as a benchmark for comparison of reported and not reported biological activity.

c. FDA

FDA-approved chemical entities were downloaded from DrugBank on Janurary 9th, 2017.[5] Molecules were excluded if they contained any of the following: < 3 carbon atoms (83), metal complexes (26), duplicates within the library (16), polymers/oligomers (5), contrast/imaging agents or dyes (4), excipients (4), sanitizers (1) and/or > 2000 amu (8). Additionally, drug cocktails were separated into individual molecules and counter cations and anions were removed to yield a final library count of 1765 molecules. A complete list of compounds can be found in the accompanying SI Excel file (OtherLibraries.xls).

2. Cheminformatic Calculations

The 20 cheminformatic parameters were adapted from Tan and co-workers, who successfully utilized the descriptors to differentiate natural products, synthetic drugs, natural product-like libraries, and drug-like libraries.[6] Two parameters were proposed to be natural product specific: number of stereocenters/molecular weight (nStereoMW) and size of largest ring (RngLg). The parameters were replaced with number of heteroatom-containing rings (HetRings) and total charge (TC), which are known to be important for RNA recognition.[1, 3a, 7] In addition, the two descriptors calculated in VCC, n-Octanol/water partition coefficient alt (ALOGPs) and Tetko's logS aqueous solubility (ALOGpS), were replaced with ChemAxon descriptors: n-Octanol/water partition coefficient (LogP) and accessible surface area (ASA).

SMILES strings for all molecules were batch processed. Using the ChemAxon Calculator Plugins, all structures were corrected to their major protonation and tautomeric states (pH = 7.4), and then the cheminformatic descriptors were evaluated using the ChemAxon Chemical Terms Evaluator (Marvin 16.4.11.0, 2016, http://www.chemaxon.com).[6] Input expressions are listed in **SI Table 2-1**.

SI Table 2-1: Cheminformatic descriptors

Category	Type	Parameter	Description	Chemical Terms Evaluator Expression
Established Medicinal Chemistry Descriptors	Lipinski's Rules	MW	Molecular Weight	mass()
		HBA	Number of Hydrogen Bond Acceptors	acceptorCount()
		HBD	Number of Hydrogen Bond Donors	donorCount()
		LogP	n-Octanol/Water Partition Coefficient	logPKLOP()
	Veber's Rules	RotB	Number of Rotatable Bonds	rotatableBondCount()
		tPSA	Topological Polar Surface Area	PSA()
	Oral Availability	LogD	n-Octanol/Water Distribution Coefficient	logD('7.4')
Additional Descriptors	Structural	N	Number of Nitrogen Atoms	atomCount('7')
		O	Number of Oxygen Atoms	atomCount('8')
		Rings	Number of Rings	ringCount()
		ArRings	Number of Aromatic Rings	aromaticRingCount()
		HetRings	Number of Heteroatom-containing Rings	heteroRingCount()
		SysRings	Number of Ring Systems	ringSystemCount()
		SysRR	Ring Complexity	ringCount()/ringSystemCount()
	Molecular Complexity	Fsp^3	Fraction of sp^3 Hybridized Carbons	count(filter('atno()==6&&connections()==4'))/atom Count('6')
		nStereo	Number of Stereocenters	chiralCenterCount()
	Molecular Recognition	ASA	Accessible Surface Area	ASA()
		relPSA	Relative Polar Surface Area	PSA()/vanDerWaalsSurfaceArea()
		TC	Total Charge	totalCharge()
		VWSA	Van der Waals Surface Area	vanDerWaalsSurfaceArea()

Possible resonance structures were not accounted for in calculations.

The R-BIND SM was tautomer and protonation checked using a second program for added rigor: Molecular Operating Environment (MOE, v2017.12) software package.[8] In general, MOE generated similar structures to ChemAxon. Of the 67 small molecules, 41 of the ligands were identical, 10 ligands had an alternate protonation state, 9 ligands had an alternate tautomer, 2 ligands had alternate protonation and tautomer states, and 6 ligands had different resonance structures.

The 20 cheminformatic parameters were re-calculated in ChemAxon utilizing the SMILES codes generated by MOE. The majority of the averages were only marginally different, including total charge (TC = 0.90 and 0.93 for MOE and ChemAxon, respectively). Similarly, the medicinal chemistry properties had negligible changes in average except for LogP (1.52 and 1.02 for MOE and ChemAxon, respectively) and LogD (0.33 and -0.11 for MOE and ChemAxon, respectively). The differences were largely attributed to two classes of compounds, benzimidazoles and quinazolines. This variation can be attributed to differences in the dominant tautomeric states chosen by the two programs and the strong dependence of solubility calculations on the tautomeric state used.[9,10]

SI Table 2-2: Average cheminformatic values for R-BIND (SM) calculated using tautomers generated in either ChemAxon or MOE.

Parameter	ChemAxon	MOE	Difference
MW	350	350	0.03
HBA	3.81	3.88	-0.07
HBD	2.43	2.40	0.03
LogP	1.02	1.52	-0.50
RotB	4.18	4.19	-0.01
tPSA	79	79	0.07
LogD	-0.11	0.33	-0.44
N	4.33	4.33	0.00
O	1.61	1.61	0.00
Rings	3.67	3.67	0.00
ArRings	2.96	2.97	-0.01
HetRings	2.16	2.16	0.00
SysRings	2.36	2.36	0.00
SysRR	1.88	1.88	0.00
Fsp3	0.27	0.27	0.00
nStereo	0.31	0.37	-0.06
ASA	574	577	-2.35
relPSA	0.18	0.18	0.00
TC	0.93	0.90	0.03
VWSA	505	507	-1.91

SI-3. Mann-Whitney U Test

All cheminformatic statistical comparisons between libraries were performed in R statistical software (v3.3.1, 2016) using an independent 2-group Mann-Whitney U Test.

a. R-BIND (SM) and R-BIND (MV)

SI Table 3-1: Statistical comparison of R-BIND (SM) and (MV) descriptors

		Means			
Type	**Parameter**	**R-BIND (SM)**	**R-BIND (MV)**	**Fold Δ**	***P***
Lipinski's Rules	MW	350	1095	-2.13	< 0.001
	HBA	3.81	12.62	-2.32	< 0.001
	HBD	2.43	10.27	-3.22	< 0.001
	LogP	1.02	1.75	-0.71	0.323
Veber's Rules	RotB	4.18	27.16	-5.50	< 0.001
	tPSA	79	320	-3.04	< 0.001
Oral Availability	LogD	-0.11	-1.94	-16.17	0.134
Structure	N	4.33	16.92	-2.91	< 0.001
	O	1.61	5.43	-2.37	< 0.001
	Rings	3.67	7.97	-1.17	< 0.001
	ArRings	2.96	6.89	-1.33	< 0.001
	HetRings	2.16	4.00	-0.85	< 0.001
	SysRings	2.36	5.30	-1.25	< 0.001
	SysRR	1.88	1.68	0.10	0.723
Molecular Complexity	Fsp3	0.27	0.40	-0.48	< 0.001
	nStereo	0.31	2.03	-5.47	< 0.001
Molecular Recognition	ASA	574	1575	-1.74	< 0.001
	relPSA	0.18	0.22	-0.23	0.005
	TC	0.93	3.27	-2.53	< 0.001
	VWSA	505	1645	-2.26	< 0.001

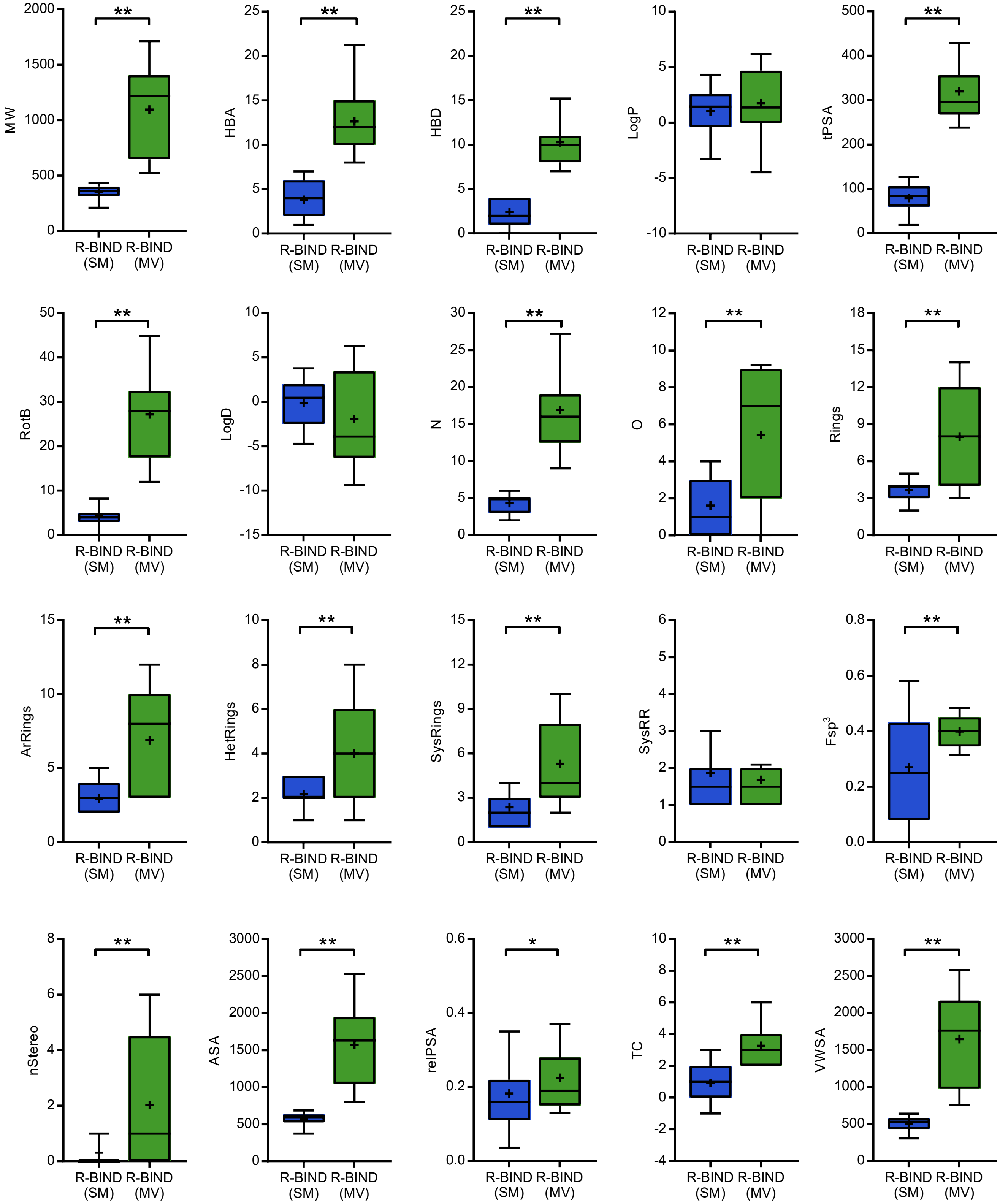
□ $P < 0.05$; ■ $P < 0.001$

SI Table 3-2: Statistical comparison of R-BIND (SM) and (MV) descriptors scaled by MW

		Means[a]			
Type	**Parameter**	**R-BIND (SM)**	**R-BIND (MV)**	**Fold Δ**	***P***
Lipinski's Rules	MW	--	--	--	--
	HBA	0.0116	0.0125	-0.0802	0.115
	HBD	0.0077	0.0115	-0.4918	0.002
	LogP	--	--	--	--
Veber's Rules	RotB	0.0113	0.0250	-1.2113	< 0.001
	tPSA	0.2473	0.3332	-0.3470	< 0.001
Oral Availability	LogD	--	--	--	--
Structure	N	0.0132	0.0170	-0.2886	0.004
	O	0.0046	0.0044	0.0345	0.611
	Rings	0.0106	0.0069	0.3431	< 0.001
	ArRings	0.0085	0.0061	0.2836	< 0.001
	HetRings	0.0063	0.0034	0.4541	< 0.001
	SysRings	0.0067	0.0047	0.3009	< 0.001
	SysRR	--	--	--	--
Molecular Complexity	Fsp3	--	--	--	--
	nStereo	0.0008	0.0018	-1.1442	0.002
Molecular Recognition	ASA	1.6559	1.4690	0.1129	< 0.001
	relPSA	--	--	--	--
	TC	0.0027	0.0038	-0.4229	0.093
	VWSA	1.4341	1.4948	-0.0423	0.011

[a] Logarithmic and fractional descriptors were not scaled

□ $P < 0.05$; ■ $P < 0.001$



SI Figure 3-1: Box-whisker plots of the 20 cheminformatic parameters for the R-BIND (SM) and (MV) ligands. The whiskers represent the 10-90th percentile of data, the boxes contain the middle 50% of the data, and the black lines and plus signs denote the medians and means, respectively. Statistically significant differences determined by the Mann Whitney U test are indicated as **P* < 0.05 and ***P* < 0.001. Abbreviations are defined in SI Table 2-1.

b. R-BIND (SM) and FDA

SI Table 3-3: Statistical comparison of R-BIND (SM) and FDA descriptors (140-590 MW, n = 1532)

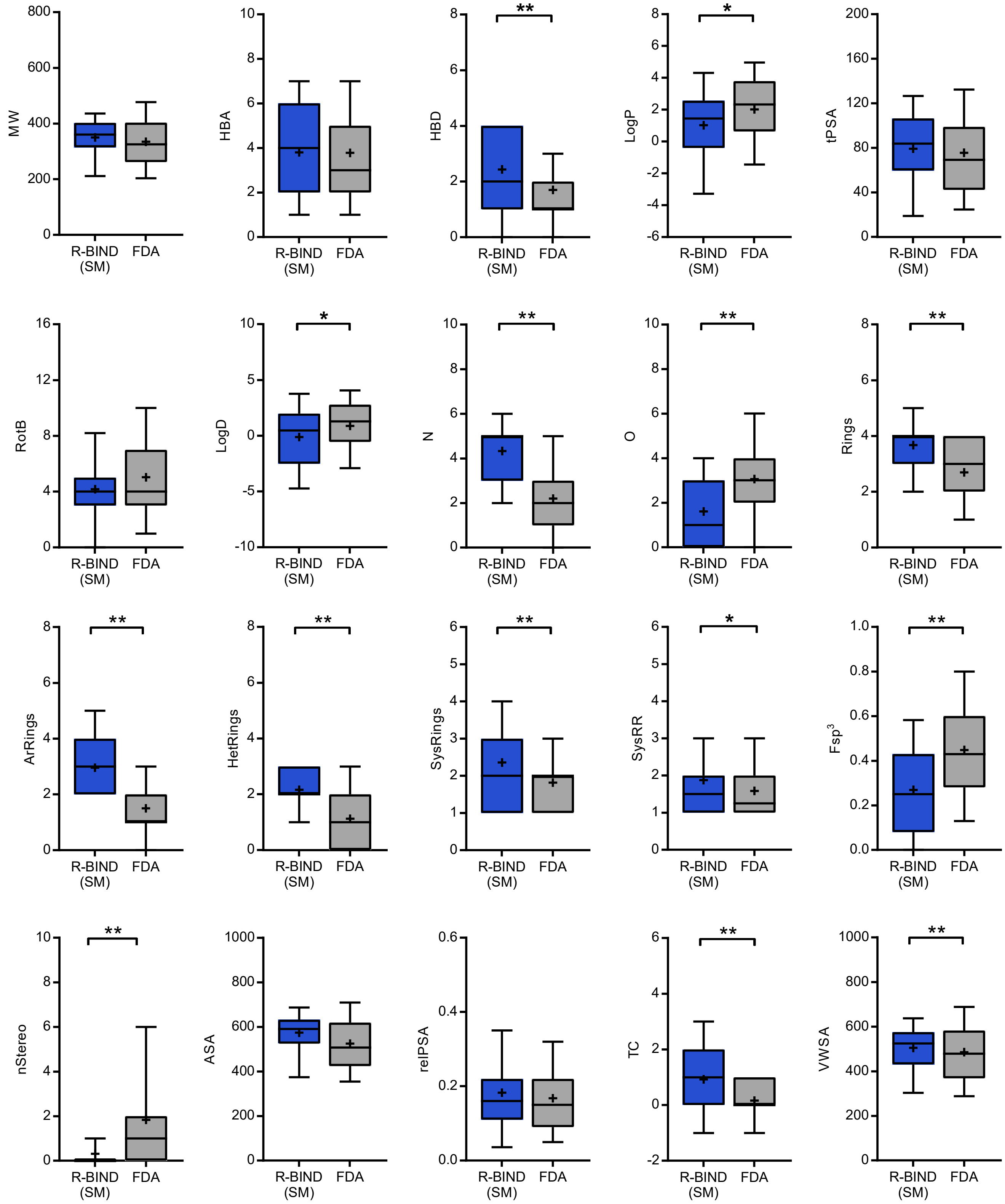
		Means			
Type	**Parameter**	**R-BIND (SM)**	**FDA**	**Fold Δ**	***P***
Lipinski's Rules	MW	350	335	0.04	0.055
	HBA	3.81	3.78	0.01	0.752
	HBD	2.43	1.70	0.30	< 0.001
	LogP	1.02	2.01	-0.97	0.002
Veber's Rules	RotB	4.18	5.03	-0.20	0.067
	tPSA	79	76	0.05	0.111
Oral Availability	LogD	-0.11	0.89	8.85	0.010
Structure	N	4.33	2.20	0.49	< 0.001
	O	1.61	3.06	-0.90	< 0.001
	Rings	3.67	2.69	0.27	< 0.001
	ArRings	2.96	1.50	0.49	< 0.001
	HetRings	2.16	1.12	0.48	< 0.001
	SysRings	2.36	1.82	0.23	< 0.001
	SysRR	1.88	1.59	0.15	0.001
Molecular Complexity	Fsp3	0.27	0.45	-0.66	< 0.001
	nStereo	0.31	1.84	-4.86	< 0.001
Molecular Recognition	ASA	574	525	0.09	< 0.001
	relPSA	0.18	0.17	0.08	0.499
	TC	0.93	0.16	0.83	< 0.001
	VWSA	505	486	0.04	0.109

□ *P* < 0.05; ■ *P* < 0.001

SI Table 3-4: Statistical comparison of R-BIND (SM) and FDA descriptors (n = 1765)

		Means			
Type	**Parameter**	**R-BIND (SM)**	**FDA**	**Fold Δ**	***P***
Lipinski's Rules	MW	350	381	-0.09	0.407
	HBA	3.81	4.49	-0.18	0.519
	HBD	2.43	2.18	0.11	0.001
	LogP	1.02	1.78	-0.74	0.008
Veber's Rules	RotB	4.18	5.84	-0.40	0.017
	tPSA	79	93	-0.17	0.658
Oral Availability	LogD	-0.11	0.60	6.33	0.036
Structure	N	4.33	2.52	0.42	< 0.001
	O	1.61	3.84	-1.38	< 0.001
	Rings	3.67	2.84	0.23	< 0.001
	ArRings	2.96	1.55	0.48	< 0.001
	HetRings	2.16	1.25	0.42	< 0.001
	SysRings	2.36	1.92	0.18	0.001
	SysRR	1.88	1.57	0.16	< 0.001
Molecular Complexity	Fsp3	0.27	0.46	-0.70	< 0.001
	nStereo	0.31	2.45	-6.81	< 0.001
Molecular Recognition	ASA	574	566	0.01	0.013
	relPSA	0.18	0.18	0.03	0.991
	TC	0.93	0.16	0.82	< 0.001
	VWSA	505	546	-0.08	0.508

□ *P* < 0.05; ■ *P* < 0.001



SI Figure 3-2: Box-whisker plots of the 20 cheminformatic parameters for the R-BIND (SM) and FDA (140-590 MW cutoff) ligands. The whiskers represent the 10-90th percentile of data, the boxes contain the middle 50% of the data, and the black lines and plus signs denote the medians and means, respectively. Statistically significant differences determined by the Mann Whitney U test are indicated as $*P < 0.05$ and $**P < 0.001$. Abbreviations are defined in SI Table 2-1.

c. R-BIND (SM) and NALDB (SM)

SI Table 3-5: Statistical comparison of R-BIND (SM) and NALDB (SM) descriptors (n = 173)

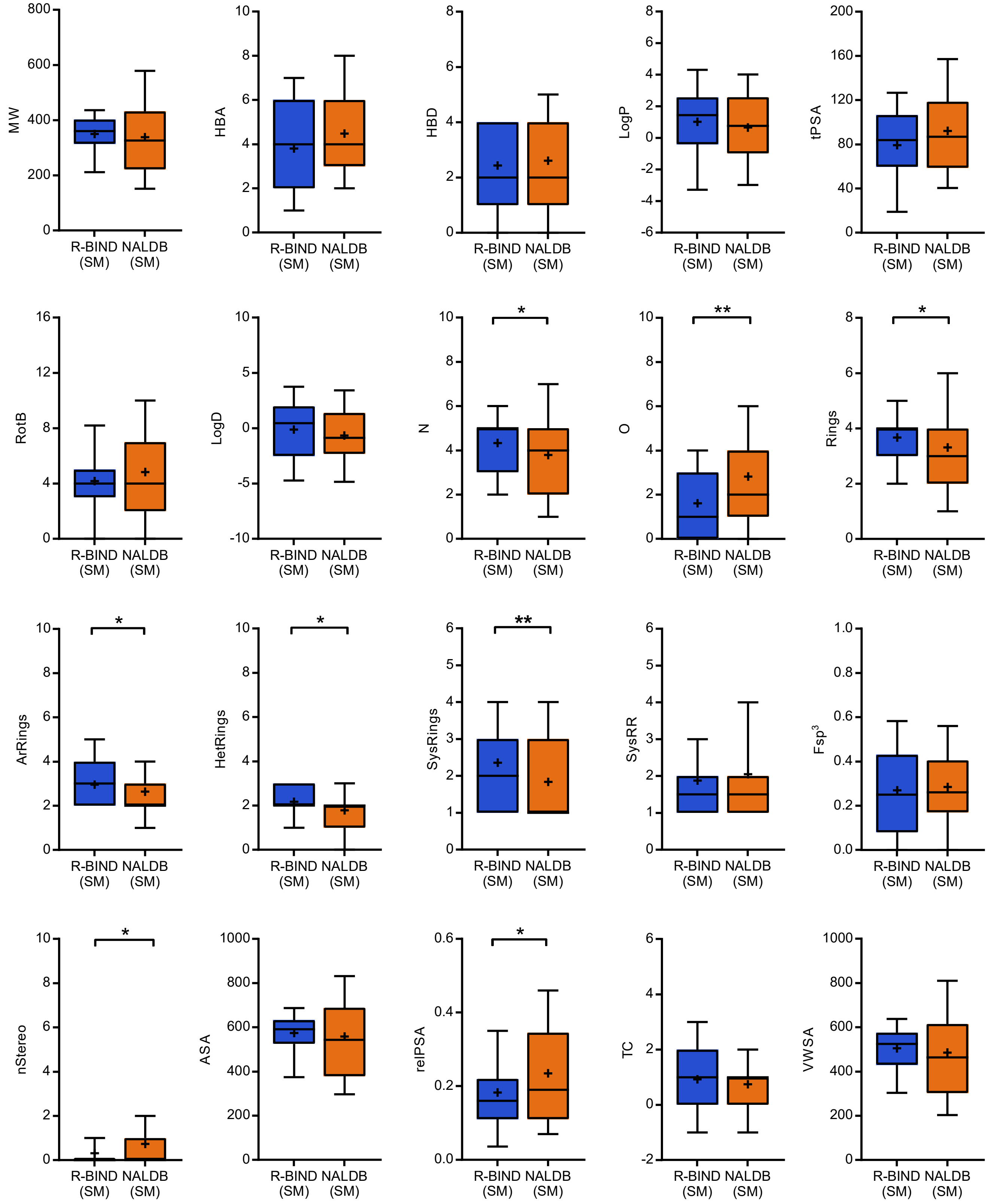
		Means			
		R-BIND	NALDB		
Type	Parameter	(SM)	(SM)	Fold Δ	*P*
Lipinski's Rules	MW	350	339	0.03	0.102
	HBA	3.81	4.48	-0.18	0.100
	HBD	2.43	2.61	-0.07	0.626
	LogP	1.02	0.66	0.35	0.242
Veber's Rules	RotB	4.18	4.84	-0.16	0.696
	tPSA	79	92	-0.16	0.189
Oral Availability	LogD	-0.11	-0.64	-4.70	0.148
Structure	N	4.33	3.79	0.12	0.012
	O	1.61	2.82	-0.75	< 0.001
	Rings	3.67	3.31	0.10	0.044
	ArRings	2.96	2.64	0.11	0.045
	HetRings	2.16	1.79	0.17	0.010
	SysRings	2.36	1.84	0.22	< 0.001
	SysRR	1.88	2.05	-0.09	0.493
Molecular Complexity	Fsp3	0.27	0.28	-0.06	0.731
	nStereo	0.31	0.73	-1.34	0.039
Molecular Recognition	ASA	574	559	0.03	0.303
	relPSA	0.18	0.23	-0.28	0.028
	TC	0.93	0.75	0.19	0.454
	VWSA	505	486	0.04	0.168

□ $P < 0.05$; ■ $P < 0.001$

SI Table 3-6: Statistical comparison of R-BIND (SM) and NALDB (SM) descriptors (140-590 MW, n = 152)

		Means			
		R-BIND-	NALDB		
Type	Parameter	(SM)	(SM)	Fold Δ	*P*
Lipinski's Rules	MW	350	336	0.04	0.081
	HBA	3.81	4.55	-0.20	0.081
	HBD	2.43	2.57	-0.06	0.762
	LogP	1.02	0.57	0.44	0.189
Veber's Rules	RotB	4.18	4.93	-0.18	0.364
	tPSA	79	93	-0.17	0.177
Oral Availability	LogD	-0.11	-0.71	-5.31	0.119
Structure	N	4.33	3.72	0.14	0.006
	O	1.61	2.89	-0.80	< 0.001
	Rings	3.67	3.26	0.11	0.038
	ArRings	2.96	2.61	0.12	0.041
	HetRings	2.16	1.72	0.21	0.005
	SysRings	2.36	1.84	0.22	0.002
	SysRR	1.88	2.00	-0.07	0.421
Molecular Complexity	Fsp3	0.27	0.30	-0.11	0.431
	nStereo	0.31	0.74	-1.35	0.026
Molecular Recognition	ASA	574	557	0.03	0.281
	relPSA	0.18	0.23	-0.24	0.036
	TC	0.93	0.67	0.27	0.300
	VWSA	505	480	0.05	0.143

□ $P < 0.05$; ■ $P < 0.001$



SI Figure 3-3: Box-whisker plots of the 20 cheminformatic parameters for the R-BIND (SM) and NALDB (SM) (without MW cutoff) ligands. The whiskers represent the 10-90th percentile of data, the boxes contain the middle 50% of the data, and the black lines and plus signs denote the medians and means, respectively. Statistically significant differences determined by the Mann Whitney U test are indicated as *P < 0.05 and **P < 0.001. Abbreviations are defined in SI Table 2-1.

SI-4. Principal Component Analysis

Each physicochemical and structural parameter was normalized to the average and standard deviation of the R-BIND, NALDB, and FDA libraries as defined in Equation (1):

$$x_{i,Norm} = \frac{x_i - \bar{x}}{s} \quad (1)$$

where x_i is the parameter value for a given molecule, and $\bar{x}$ and s represent the mean and standard deviation of a parameter, respectively, for the combined libraries. Principal component analysis (PCA) was performed on the normalized data using the Microsoft Excel add-in, XLSTAT (v18.07.40123, 2017, Addinsoft). The analysis was run as a Spearman PCA, and the factor scores were used for visualization and nearest neighbor clustering analysis.

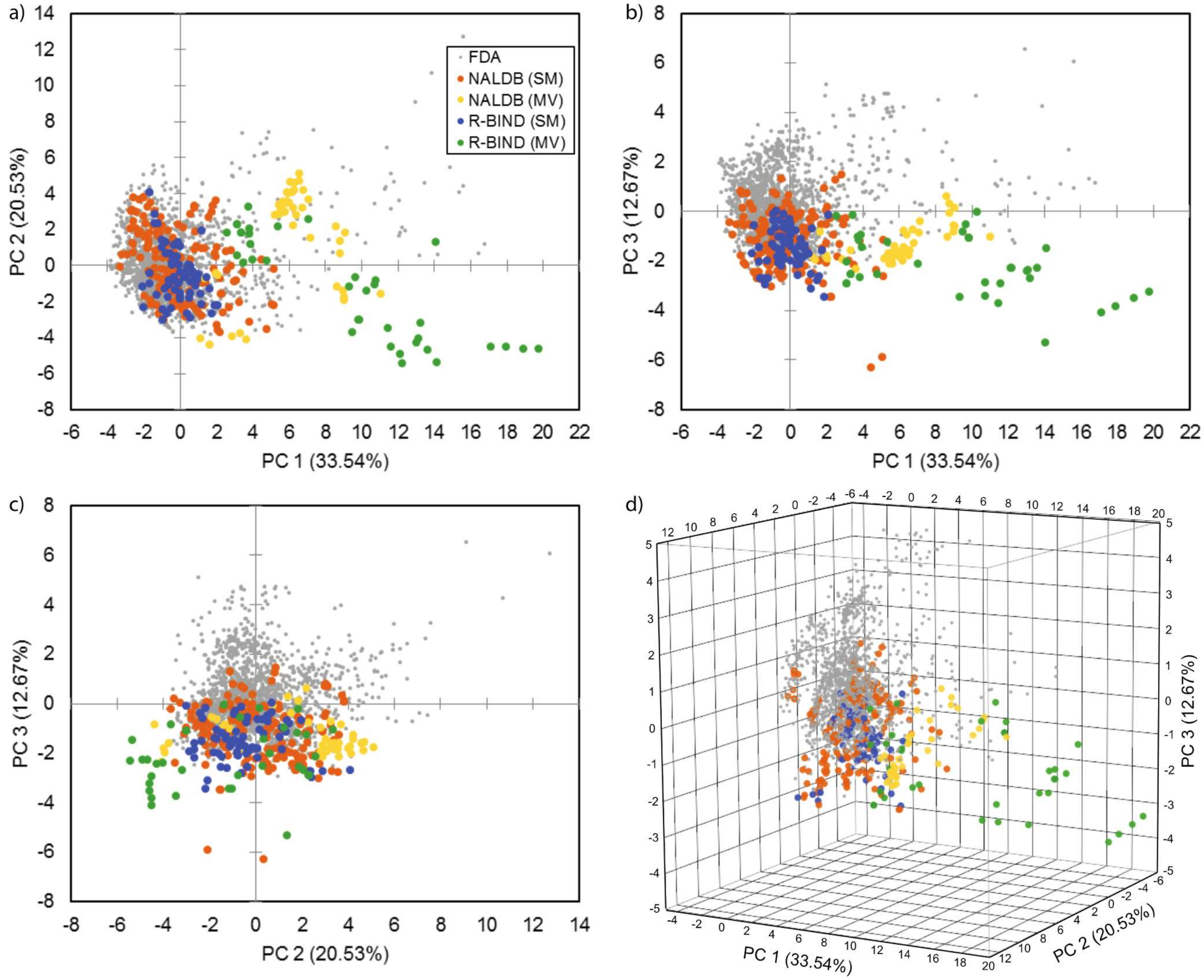
SI Table 4-1: Eigenvalues of each principal component

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9	PC 10
Eigenvalue	6.708	4.106	2.533	1.702	1.584	0.746	0.645	0.499	0.309	0.251
Variability (%)	33.538	20.528	12.667	8.508	7.921	3.729	3.226	2.495	1.543	1.257
Cumulative %	33.538	54.066	66.733	75.241	83.162	86.891	90.117	92.612	94.155	95.412

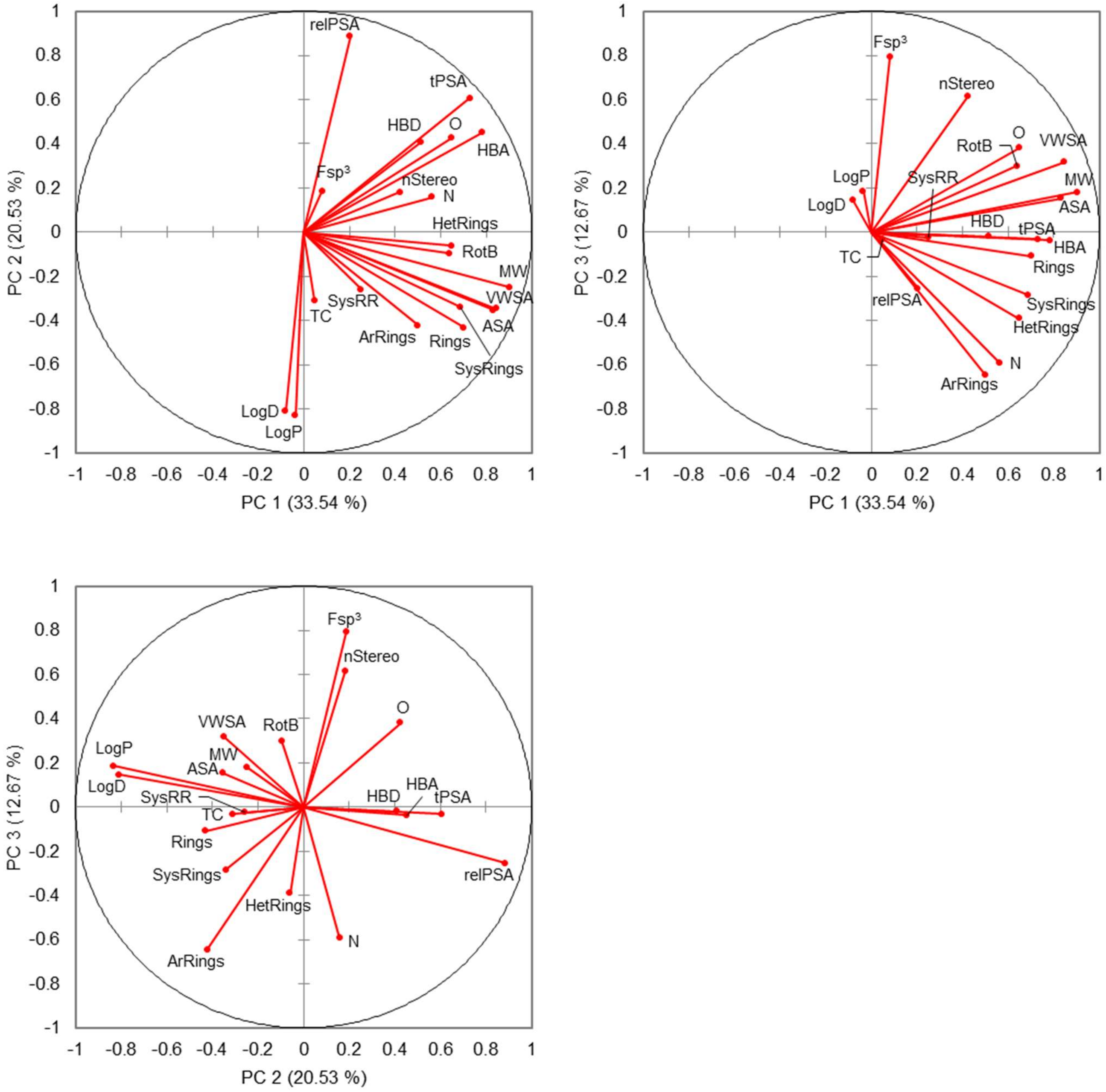
	PC 11	PC 12	PC 13	PC 14	PC 15	PC 16	PC 17	PC 18	PC 19	PC 20
Eigenvalue	0.209	0.154	0.131	0.103	0.094	0.079	0.054	0.045	0.028	0.020
Variability (%)	1.043	0.772	0.654	0.515	0.472	0.397	0.272	0.227	0.138	0.100
Cumulative %	96.455	97.227	97.881	98.395	98.867	99.264	99.535	99.763	99.900	100.000

SI Table 4-2: Percent contributions of each parameter for each principal component

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9	PC 10
MW	12.149	1.501	1.293	0.204	0.253	0.238	0.923	0.166	0.571	7.266
HBA	9.121	4.958	0.056	0.256	4.433	0.226	0.010	4.026	5.503	0.887
HBD	3.963	4.044	0.014	0.490	12.210	25.206	16.033	0.107	3.481	0.624
LogP	0.019	16.842	1.360	2.293	3.433	5.903	5.877	6.022	0.307	3.285
RotB	6.053	0.224	3.553	17.263	4.119	0.023	6.067	0.042	1.836	18.456
tPSA	7.919	8.991	0.041	0.666	0.936	3.253	0.457	1.552	1.509	0.085
LogD	0.096	15.924	0.858	0.686	4.613	3.139	11.149	16.138	2.755	0.392
N	4.725	0.611	13.917	0.024	6.039	1.320	0.054	22.124	9.869	0.165
O	6.266	4.418	5.717	0.252	7.225	0.095	0.548	3.897	19.024	2.574
Rings	7.346	4.532	0.457	12.696	1.206	0.185	0.343	2.075	0.000	7.526
ArRings	3.729	4.335	16.393	0.277	0.102	1.474	0.853	6.212	1.031	10.319
HetRings	6.269	0.089	6.000	7.795	0.001	20.167	0.456	6.241	6.947	13.366
SysRings	7.024	2.804	3.173	1.566	0.134	12.470	18.667	8.266	0.017	6.654
SysRR	0.940	1.637	0.024	37.201	1.822	10.696	14.683	0.851	0.429	0.420
Fsp^3	0.098	0.848	24.794	2.267	5.095	10.840	0.001	13.305	0.037	1.687
nStereo	2.657	0.806	14.850	9.727	0.072	0.450	13.719	6.601	11.013	20.141
ASA	10.328	3.040	0.945	4.197	0.482	0.065	6.322	0.019	0.059	0.221
relPSA	0.616	19.134	2.572	0.292	1.783	2.206	1.507	0.411	0.464	0.051
TC	0.033	2.339	0.041	1.592	45.397	1.967	0.285	1.925	34.079	0.221
VWSA	10.649	2.921	3.942	0.256	0.644	0.078	2.046	0.020	1.070	5.661



SI Figure 4-1: Extended principal component analysis plots based on the cheminformatic parameters calculated for the R-BIND, NALDB, and FDA libraries. a) PCA plot of PC 1 versus PC 2. b) PCA plot of PC 1 versus PC 3. c) PCA plot of PC 2 versus PC 3. d) PCA plot of PC 1-3. The principal component and subsequent percent contribution is indicated on each axis.



SI Figure 4-2: Loading plots for the first three principal components. a) Loading plot of PC 1 versus PC 2. b) Loading plot of PC 1 versus PC 3. c) Loading plot of PC 2 versus PC 3. The magnitude and direction of the vector indicates the contribution of that parameter to the component. The percent contribution of each principal component is indicated.

SI-5. Nearest Neighbor Clustering Analysis

The nearest neighbor (NN) clustering analysis is an analog to the k nearest neighbor (k-NN) classification algorithm. Due to the limited number of data points in each dataset, we applied a 1-NN algorithm to study overlap. The algorithm follows the procedure described below:

(1) For each dataset, the multi-dimensional distances in space are calculated for each pair of data points; (2) The average nearest neighbor distance for this dataset is calculated; (3) Each point is mapped as a multi-dimensional sphere with a radius of the averaged nearest neighbor distance; (4) Data points from other datasets are mapped to the same space and overlaps are counted for the populated regions generated in step (3).

To count the overlap within a dataset (i.e. FDA molecules in FDA), a data point is considered to be within the cluster if it overlapped with regions populated by other data points in the same dataset. To count overlap between datasets (i.e. FDA molecules in R-BIND (SM)), a data point is considered to be within another library's cluster if it overlapped with the NN defined cluster.

SI Table 5-1: Nearest neighbor quantification of principal component analysis in three dimensions (67% of variance)

	Library Cluster				
	FDA	**R-BIND (SM)**	**R-BIND (MV)**	**NALDB (SM)**	**NALDB (MV)**
FDA molecules in…	1192	250	6	440	1
R-BIND (SM) molecules in…	36	42	0	21	2
R-BIND (MV) molecules in…	1	0	21	0	0
NALDB (SM) molecules in…	98	33	2	104	0
NALDB (MV) molecules in…	1	1	1	0	27

SI Table 5-2: Nearest neighbor quantification of principal component analysis in ten dimensions (95% of variance)

	Library Cluster				
	FDA	**R-BIND (SM)**	**R-BIND (MV)**	**NALDB (SM)**	**NALDB (MV)**
FDA molecules in…	1021	167	0	157	0
R-BIND (SM) molecules in…	11	36	0	9	0
R-BIND (MV) molecules in…	0	0	27	0	0
NALDB (SM) molecules in…	54	33	0	96	0
NALDB (MV) molecules in…	0	0	1	0	28

SI-6. Principal Moments of Inertia Calculations

Ligands in the NALDB (MV) and R-BIND (MV) libraries were excluded from this analysis to avoid potential bias in modeling larger molecular weight ligands.[11] Similarly, calculations were performed on the molecular weight restricted NALDB (SM) and FDA libraries (140-590 amu) to avoid modeling bias and to compare analogous libraries.

Low energy conformations of each molecule, using the protonation- and tautomer-corrected SMILES strings (SI-2), were calculated using the Conformation Search algorithm in the Molecular Operating Environment (MOE, v2017.12) software package.[8] The Conformation Search function was performed using the stochastic method with the MMFF94 force field and generalized Born solvation model. The input for each parameter is listed in **SI Table 6-1**, and the following options were checked: calculate force field partial charges and hydrogens.

SI Table 6-1: Parameters for conformation search

Parameter	Input
Rejection limit	100
Iteration limit	10000
RMS gradient	0.005
MM iteration limit	500
RMSD limit	0.15
Energy window	3
Conformation limit	10000

The 3 kcal/mol energy window was selected to survey biologically-relevant conformation space[12] and to obtain a representative population of conformers at equilibrium (> 99%) as described by Equation (3).

$$\frac{N_1}{N_0} = e^{-\Delta E/RT} \tag{3}$$

where N_1/N_0 is the ratio of the number of molecules in the relative energy states, ΔE is the energy difference between N_0 and N_1 (3 kcal/mol), R is the ideal gas constant (0.00198588 kcal/K mol), and T is the temperature (298 K).

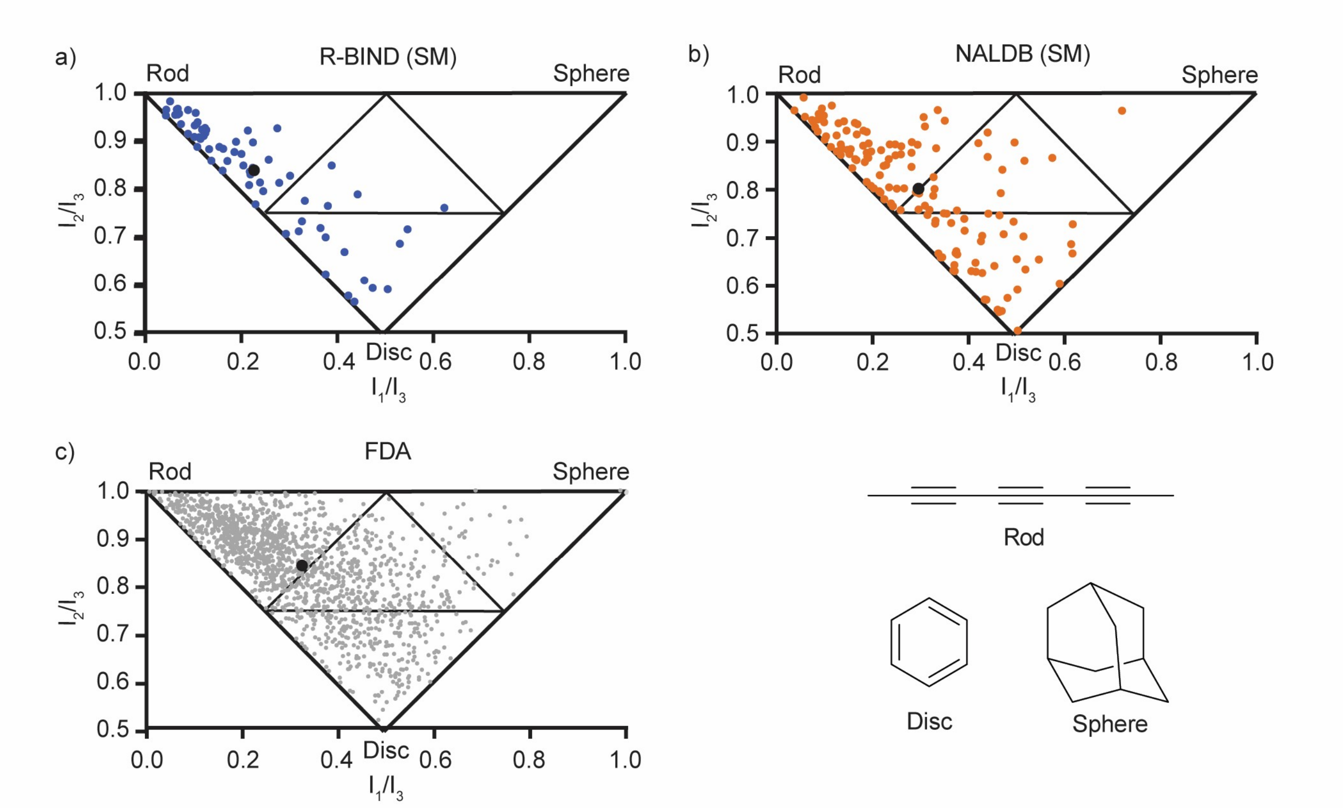
After the conformational search was complete, the normalized principal moment of inertia descriptors, *npr1* (I_1/I_3) and *npr2* (I_2/I_3), were computed for each conformer in MOE. The Boltzmann weighted average for *npr1* and *npr2* of each molecule was calculated by using Equation (4).

$$A_i = \frac{\sum_i A e^{-\frac{\mathrm{E}_i}{\mathrm{k_B T}}}}{\sum_i e^{-\frac{\mathrm{E}_i}{\mathrm{k_B T}}}} \tag{4}$$

where *A* is the calculated *npr1* or *npr2* value of the conformation, E_i is the relative energy of the conformation (kcal/mol), k_B is the Boltzmann constant (0.001986 kcal/(mol*K)), and T is the temperature (298 K). The resulting coordinates were plotted on a triangular graph where the vertices represent rod- (0,1), sphere- (1,1), or disc-like (0.5, 0.5) shape.

SI Table 6-2: Average normalized principal moment of inertia for each library

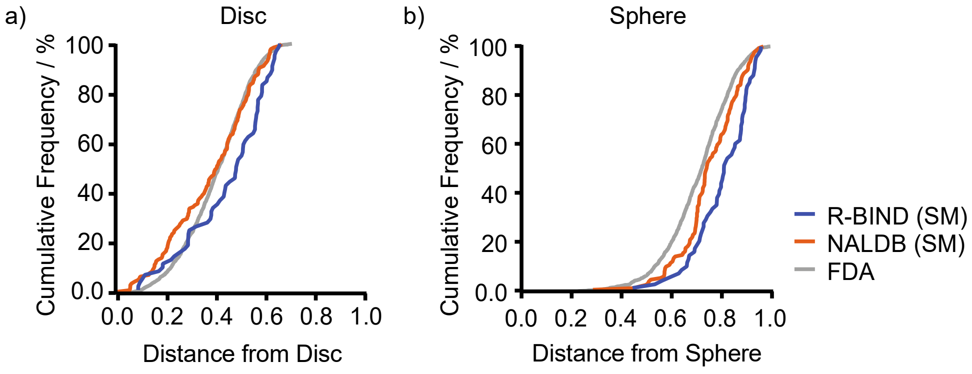
Library	I_1/I_3	I_2/I_3
R-BIND (SM)	0.23	0.84
NALDB (SM)	0.29	0.80
FDA	0.32	0.84



SI Figure 6-1: Triangle plots of normalized principal moments of inertia for the a) R-BIND (SM), b) NALDB (SM) (140-590 amu) and c) FDA libraries (140-590 amu). The four sub-triangles represent the general shapes of rod, hybrid, disc, and sphere. Each colored dot represents the Boltzmann average of a molecule using conformations within 3 kcal/mol of the lowest energy conformer. The black dot represents the average shape of the library.

a. Cumulative Distribution

The Euclidean distance of each small molecule coordinate from the rod (0,1), sphere (1,1), and disc (0.5, 0.5) vertices was calculated. The distances for each library were ordered from smallest to largest, and the cumulative frequency for each distance was calculated. The cumulative distribution graphs were generated using GraphPad Prism (version 7.02 for Windows, GraphPad Software, La Jolla California USA, www.graphpad.com).



SI Figure 6-2: Cumulative distribution of the distance from the a) disc and b) sphere vertices for R-BIND (SM), NALDB (SM), and FDA libraries (140-590 amu). Cumulative distribution of the distance from the rod vertex is in Figure 4.

b. Kolmogorov-Smirnov Test

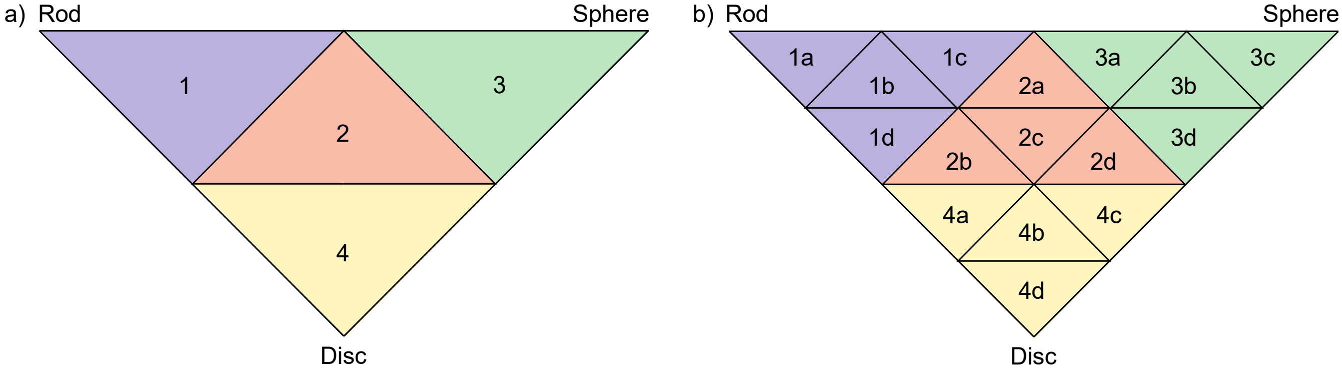
The rod, disc, and sphere distributions for each library were compared using a two-sided, two-sample, non-parametric Kolmogorov-Smirnov test. The test was run in R statistical software (v3.3.1, 2016) using R Commander (Rcmdr, v2.3-2, 2017).

SI Table 6-3: Statistical comparisons of the distributions from the rod, disc, and sphere vertices for each pair of libraries

	***P* values**		
Shape	**R-BIND (SM) / FDA**	**R-BIND (SM) / NALDB (SM)**	**FDA / NALDB (SM)**
Rod	< 0.001	0.009	0.442
Disc	< 0.001	0.016	0.025
Sphere	< 0.001	0.007	< 0.001

c. Cell-Based Partitioning

Cell-based partitioning was performed in Python (Python Language Reference v2.7, http://www.python.org). The principal moments of inertia triangle was defined by three lines: y = 1; y = -x + 1; and y = x. The triangle was then partitioned into four or sixteen isosceles triangles of equal size by defining the slopes and area of each sub-triangle. The Boltzmann weighted average coordinate of each small molecule was rounded to three significant figures, and if the coordinate fell within a partition, the script returned the identity of the triangle in which the value was located.



SI Figure 6-3: Principal moments of inertia triangle partitions. The identity and locations are listed for a) four and b) sixteen triangle partitions.

SI Table 6-4: Small molecule counts for each library in the four triangle partitions

Library	**1**	**2**	**3**	**4**	**Total**
R-BIND (SM)	48	5	0	14	67
NALDB (SM)	82	20	1	44	147
FDA	857	369	33	273	1532

SI Table 6-5: Small molecule counts for each library in the sixteen triangle partitions

	Rod				**Hybrid**				**Sphere**				**Disk**				
Library	**1a**	**1b**	**1c**	**1d**	**2a**	**2b**	**2c**	**2d**	**3a**	**3b**	**3c**	**3d**	**4a**	**4b**	**4c**	**4d**	**Total**
R-BIND (SM)	26	8	0	14	0	4	0	1	0	0	0	0	7	1	1	5	67
NALDB (SM)	30	19	3	30	3	12	5	0	0	1	0	0	18	12	3	11	147
FDA	240	308	57	252	45	188	87	49	14	8	1	10	111	75	35	52	1532

SI-7. References for SI

[1] J. R. Thomas, P. J. Hergenrother, *Chem. Rev.* **2008**, *126*, 224.

[2] E. Jankowsky, M. E. Harris, *Nat. Rev. Mol. Cell Biol.* **2015**, *16*, 533.

[3] a) F. Aboul-ela, *Future Med. Chem.* **2010**, *2*, 93; b) M. D. Disney, A. J. Angelbello, *Acc. Chem. Res.* **2016**, *49*, 2698.

[4] S. Kumar Mishra, A. Kumar, *Database* **2016**, *2016*.

[5] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J. Woolsey, *Nucleic Acids Res.* **2006**, *34*, D668.

[6] T. A. Wenderski, C. F. Stratton, R. A. Bauer, F. Kopp, D. S. Tan, *Methods Mol. Biol.* **2015**, *1263*, 225.

[7] a) W. D. Wilson, K. Li, *Curr. Med. Chem.* **2000**, *7*, 73; b) T. Hermann, *Wiley Interdiscip. Rev.: RNA* **2016**.

[8] Molecular Operating Environment (MOE), 2013.08; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2017.

[9] Y. C. Martin, *J. Comput.-Aided Mol. Des.* **2009**, *23*, 693.

[10] S. W. Muchmore, J. J. Edmunds, K. D. Stewart, P. J. Hajduk, *J. Med. Chem.* **2010**, *53*, 4830.

[11] M. Wirth, W. H. B. Sauer, *Mol. Inf.* **2011**, *30*, 677.

[12] a) F. Kopp, C. F. Stratton, L. B. Akella, D. S. Tan, *Nat. Chem. Biol.* **2012**, *8*, 358; b) J. Bostrom, P. O. Norrby, T. Liljefors, *J. Comput.-Aided Mol. Des.* **1998**, *12*, 383.

8. Extended References

[14] J. Palacino, S. E. Swalley, C. Song, A. K. Cheung, L. Shu, X. L. Zhang, M. Van Hoosear, Y. Shin, D. N. Chin, C. G. Keller, M. Beibel, N. A. Renaud, T. M. Smith, M. Salcius, X. Y. Shi, M. Hild, R. Servais, M. Jain, L. Deng, C. Bullock, M. McLellan, S. Schuierer, L. Murphy, M. J. J. Blommers, C. Blaustein, F. Berenshteyn, A. Lacoste, J. R. Thomas, G. Roma, G. A. Michaud, B. S. Tseng, J. A. Porter, V. E. Myer, J. A. Tallarico, L. G. Hamann, D. Curtis, M. C. Fishman, W. F. Dietrich, N. A. Dales, R. Sivasankaran, *Nat. Chem. Biol.* **2015**, *11*, 511.

[19] R. Santos, O. Ursu, A. Gaulton, A. P. Bento, R. S. Donadi, C. G. Bologa, A. Karlsson, B. Al-Lazikani, A. Hersey, T. I. Oprea, J. P. Overington, *Nat. Rev. Drug Discovery* **2017**, *16*, 19.